Report - Assignment 2

BALAMURUGAN Siva Sathya Pradha(50169073)
GOPALSAMY Rahul (50163719)
SUNDARA RAMAN Bhavani(50169253)

1 Problem 1: Experiment with Gaussian Discriminators

1.1 Introduction

1.1.1 Linear Discriminant Analysis (LDA)

LDA was performed on sample.pickle data set. There were 150 data points with two predictors x_1&x_2 in training data. Each data point belongs to k=5 classes 1,2,3,4,5. LDA attempts to approximate Bayes classifier. Given the distribution of predictors within each class, it estimate the probability of new observation belonging to a particular class. The decision boundary between classes is linear.

The assumptions in LDA:

- Each observation is drawn from Gaussian distribution of its respective class.
- Homogeneous variance-covariance matrices.

1.1.2 Quadratic Discriminant Analysis (QDA)

QDA is similar to LDA except that it considers individual variance and covariance matrix for each class. The resulting decision boundary between classes is non-linear.

The assumptions in QDA:

- Each observation is drawn from Gaussian distribution of its respective class.
- Heterogeneous variance-covariance matrices.

1.2 Accuracy

	Accuracy
LDA	97%
QDA	96%

1.3 Error Analysis & Discussion

The difference between LDA and QDA can be seen as **Bias-Variance trade-off**. The LDA has lesser variance when compared to QDA. As the data set is smaller in size it is necessary to **minimize the variance**. This is to be done to avoid potential over-fitting of data. Also, The two predictors are not correlated so our assumption of common variance holds true. Thus, LDA performs slightly better than the QDA.

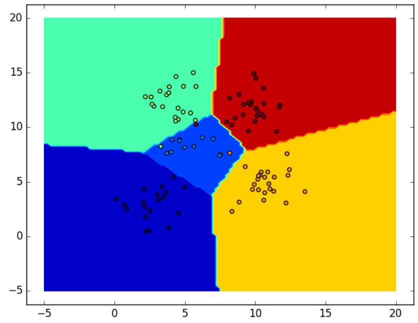


Figure 1: LDA: The decision boundary is linear. More Bias.

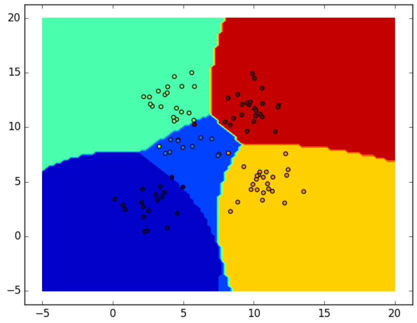


Figure 2: QDA: The decision boundary is non-linear. More Variance

2 Problem 2: Experiment with Linear Regression

2.1 Linear Regression

The linear regression is used for predicting the quantitative response. This model makes a strong assumption about the linearity between predictors. A medical data set given in diabetes.pickle was used to fit the linear model. There were a total of 64 predictors in the data set with 242 observations. The model was to learn the relationship between the patients diabetic condition and his corresponding measurements.

The Root Mean Square Error (RMSE) is used to assess the fit of the models. The lower value of RMSE indicates better prediction from the linear model. RMSE quantifies absolute fit whereas R^2 quantifies the relative fit.
Here two models are considered

- RMSE without intercept
- RMSE with intercept

2.2 Accuracy

	Without Intercept	With Intercept
RMSE on Train Data	138.20074835	46.7670855937
RMSE on Test Data	326.7649943893434	60.89203709368397

2.3 Discussion

Intercept in the linear regression is the mean of response variable when all the predictors are set to zero. Thus, intersect can be regarded as bias. Adding the bias to the model reduces the variance. When the bias is not added the predictors have to account for more variance which results in over-fitting of the data.

3 Problem 3: Experiment with Ridge Regression

3.1 Ridge Regression

Ridge regression is similar to Ordinary Least Square regression (OLS), expect that it has a tuning parameter λ to impose penalty in the learned weights. When $\lambda = 0$ Ridge regression behaves just like OLS. The Ridge regression uses l_2 norm which forces the regression coefficient estimate to approach zero as the value of $\lambda \to \infty$. The coefficients of the predictors depends upon the value of λ.

3.2 Result

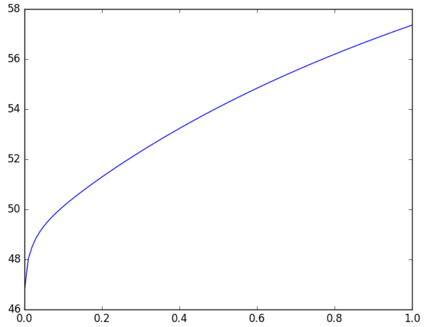


Figure 3: Train error: RMSE vs λ

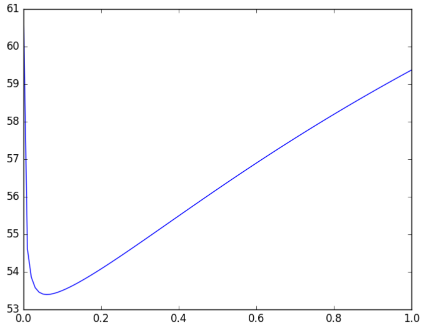


Figure 4: Test error: RMSE vs λ

- The graph in Figure 3 shows an increase in RMSE in training data as λ increase. The higher value of λ will force the coefficients of weights to be minimum resulting in more RMSE.
- The graph in Figure 4 shows at first, RMSE in testing data reduces with increase in the λ. After a critical value the RMSE again start to increase. This is because the higher penalty by λ will force the model to ignore crucial elements of the data.

The value of λ with lowest train error is its optimal value.

3.3 Discussion

Data inherently contains noise which should not be learned by the model. While training, this noise should not be taken into account. A over-fitted model to the training set results in training error value close to zero, but the model would perform poorly in the test data. This is because of the model's failure to generalize. In ridge Regression λ is the tuning parameter. The value of λ decides the Bias-Variance trade off. The higher value of λ imposes higher penalty there by forcing the weights of predictors to be approaching to zero, resulting in increased Bias. When λ is low the variance increases, Ridge regression behaves similar to OLS.

4 Problem 4: Using Gradient Decent for Ridge Regression

4.1 Gradient Decent

Gradient decent can be used in learning weights of the Ridge regression instead of minimizing regularized squared error loss. This reduces the computation involved in learning weights.

4.2 Result

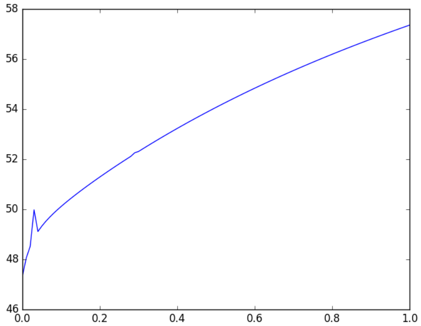


Figure 5: Train error: RMSE vs λ

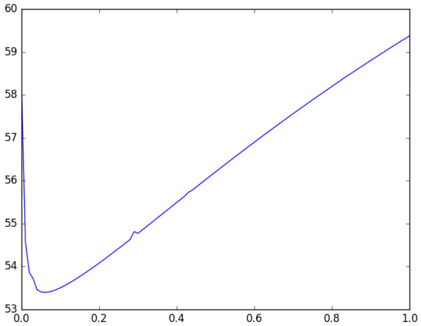


Figure 6: Test error: RMSE vs λ

4.3 Discussion

This graph is similar to the graph in problem 3. There is sight increase in RMSE value. This can be seen as a trade off between the computation and accuracy.

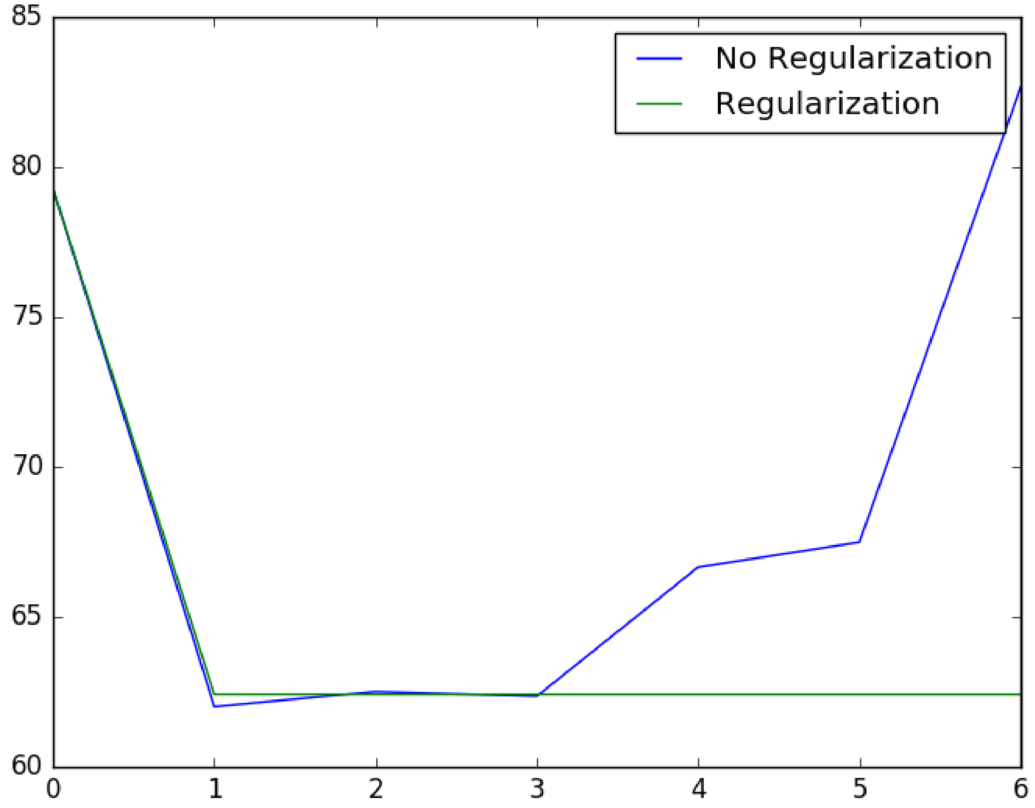
5 Problem 5: Nonlinear Regression

Linear regression can be used to produce nonlinear boundaries by addition of quadratic terms to the model. Here, We take a particular predictor x and use a vector of nonlinear terms to study the effect of it on RMSE of the test data.
Two models are considered.

- $\lambda = 0$. This results in analyzing the affect of nonlinear attributes in OLS.
- Optimal value of λ from the previous problem. This results in analyzing the effect of nonlinear terms in Ridge Regression.

As the nonlinear terms produce nonlinear boundary the flexibility (variance) of the model will be high. Therefore, We hypothesize that the nonlinear model with regularization would perform better than the nonlinear model without regularization, as the p value increases. This is because the flexibility in the non linear model would be balanced by adding penalties (bias) to the learned weight.

5.1 Result



5.2 Discussion

The resulting graph shows the RMSE value with respect to a predictor x as a vector of p $\{1, x, x^2, ...x^p\}$ attributes. For a model without a regularization (OLS Regression), the RMSE value decreases up to $p = 3$. After that increase in p value causes the RMSE value to increase. This is because the model tends to overfit on the training data resulting in higher error in testing data. The flexibility of the model will increase with increase in p, causing the training error to reduce, but test error will increase. In this model, we would prefer p=1 as this results in lowest RMSE value.

The model with regularization offsets the increase in model flexibility by forcing the weights towards zero. Thus, Bias increases resulting in better RMSE. The RMSE almsot stays stable with increase in p.In this model, we would prefer p=1. According to **Occam's Razor** a simple model should be used.

6 Problem 6: Interpreting Results

The Linear regression models is type of parametric method. The regression model in our case should predict the diabetic condition of the patient based on the a set of predictors. We are not concerned about the form of learned function as along as the its has high prediction accuracy.

In order to evaluate a performance of a statistical model, Root mean square Error is used to assess its fit. The RMSE is small for model with good prediction and is high for model with poor prediction. There are two types of error : 1) Train error 2) Test error. Train error indicates how well the model has learned from the training data. Test error indicates how well the model predicts an unseen observation. There is no guarantee that the model with small train error will also have small test error. We are concerned about the performance of the model in terms of test error.

Requirements from the model for our problem.

- High Prediction Accuracy.
- less Variance.
- Low test error.
- simpler model.

6.1 Comparisons

	Advantages	Disadvantages
OLS Regression	When used with intercept results in lower RMSE value than the nonlinear model.	Does not have the penalty term on the learned weight from the model.
Ridge Regression	Has the lowest RMSE value for our problem	Evaluating optimal value of λ using cross validation can be a problem as the size of data set increases.
Gradient descent	Avoids the computation of $(X^T X)^{-1}$	Slight increase in RMSE value compared to the squared error loss.
Nonlinear OLS Regression	Gives the flexibility to the model	Tends to overfit the data with increase in p
Nonlinear Ridge Regression	Offsets the flexibility in the model by introducing penalty in the learned weight.	The value of RMSE does not significantly decrease with increase in p after a level. The increase in attributes results in unnecessary complicated model.

The recommendation is to use the **Ridge regression model**. This model would result in the least test error. The optimal value of λ can be estimated by cross validation. Generalization is important in terms of model's prediction accuracy of an unseen observation.